Deep RegulAtory GenOmic Neural Networks - DragoNN

Practice **LESS** Deep Learning

Learn - **E**xperiment - **S**hare - **S**eek

Barathi Ganesh HB

Centre for Excellence in Computational Engineering and Networking (CEN)
Amrita School of Engineering, Coimbatore
Amrita Vishwa Vidyapeetham, India
email: barathiganesh.hb@gmail.com

Outline

Genomics

DNA sequencing

Nucleotides

Property of Regulatory Sequence

DragoNN

Simulations with DragoNN

Representation of motifs

Classification with DragoNN

Genomics

- An interdisciplinary field of *science* within the field of *molecular biology*.
- Aims at the collective characterization and quantification of genes.
- Direct the production of proteins with the assistance of enzymes and messenger molecules.
- Uses high throughput **DNA sequencing** and bioinformatics to assemble, and analyze the function and structure of entire genomes.

DNA sequencing

- DNA sequencing is the process of determining the precise order of **nucleotides** within a DNA molecule.
- It includes any method or technology that is used to determine the order of the four bases - adenine (A), guanine (G), cytosine (C), and thymine (T) in a strand of DNA.

Nucleotides

- Organic molecules composed of three sub unit molecules: a nitrogenous base, a five-carbon sugar (ribose or deoxyribose), and at least one phosphate group.
- **Ribose** - is a carbohydrate (simple sugar)
- **Deoxyribose** - deoxy sugar - derived from the sugar ribose by loss of an oxygen atom.
- **Nitrogenous base** - Adenine (A), Guanine (G), Thymine (T), Cytosine (C), Uracil (U)
- **A phosphate** (PO34) is an inorganic chemical and a salt-forming anion of phosphoric acid.

Ribose, Deoxyribose and Phosphate

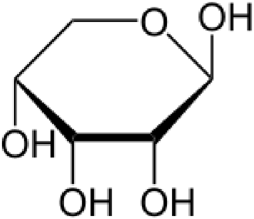


Figure: Ribose

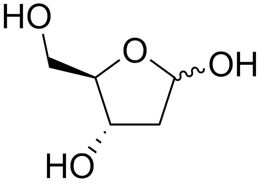


Figure: Deoxyribose

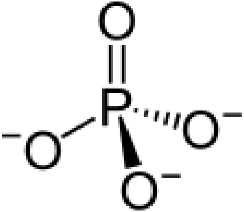
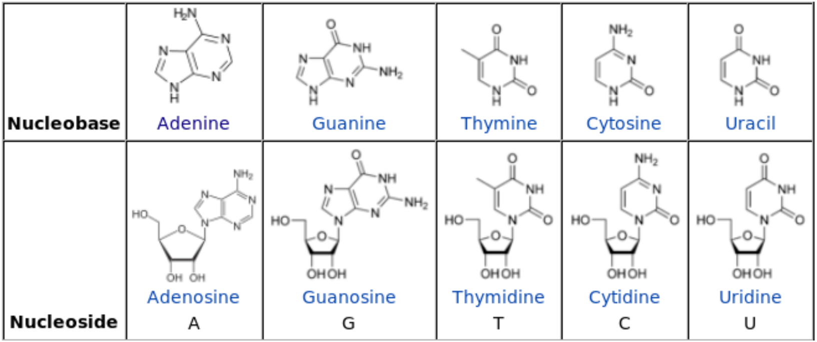


Figure: Phosphate

Nitrogenous bases



Nucleobase	Adenine	Guanine	Thymine	Cytosine	Uracil
Nucleoside	Adenosine A	Guanosine G	Thymidine T	Cytidine C	Uridine U

Figure: Nitrogenous bases

Four bases in a strand of DNA

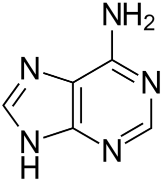


Figure: adenine

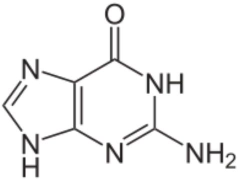


Figure: guanine

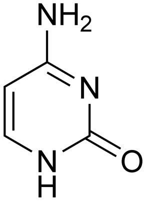


Figure: cytosine

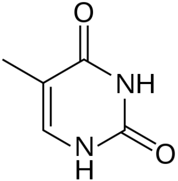
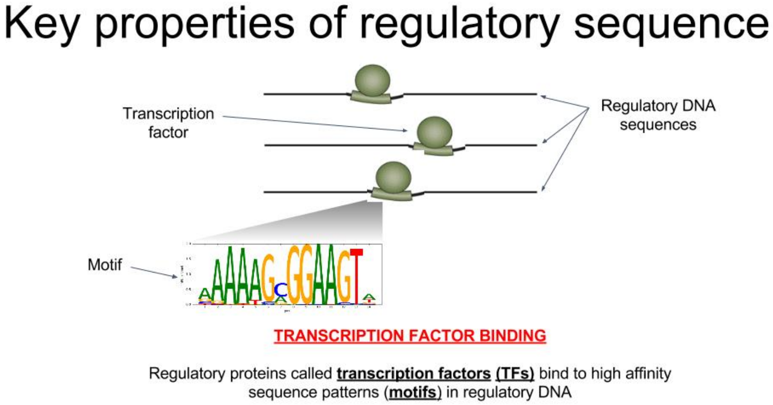


Figure: thymine

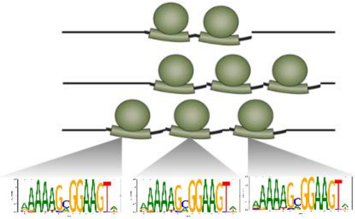


Key properties of regulatory sequence

TRANSCRIPTION FACTOR BINDING

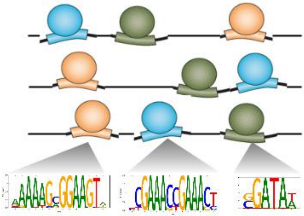
Regulatory proteins called **transcription factors (TFs)** bind to high affinity sequence patterns (**motifs**) in regulatory DNA

Figure: Nuc. level importance (height of letter) shows coordination of multiple point binding events



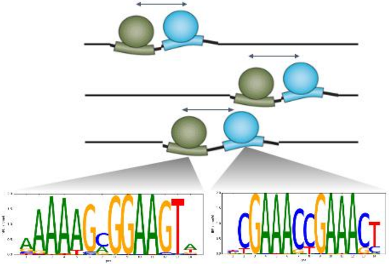
HOMOTYPIC MOTIF DENSITY

Regulatory sequences often contain **more than one binding instance** of a TF resulting in **homotypic clusters of motifs of the same TF**



HETEROTYPIC MOTIF COMBINATIONS

Regulatory sequences often bound by **combinations of TFs** resulting in **heterotypic clusters of motifs of different TFs**



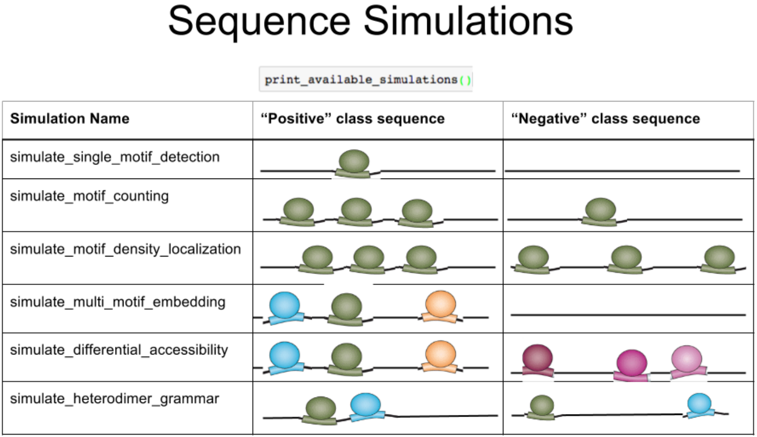
SPATIAL GRAMMARS OF HETEROTYPIC MOTIF COMBINATIONS

Regulatory sequences are often bound by **combinations of TFs** with specific **spatial and positional constraints** resulting in distinct **motif grammars**

Deep RegulAtory GenOmic Neural Networks

- A toolkit to teach and learn about deep learning for genomics.
- Enables computational biologists working on genomics problems to get started with deep learning.
- deep learning practitioners to get started with applications in genomics.
- Software for model development, model interpretation, and DNA sequence simulations.

DNA sequence simulations

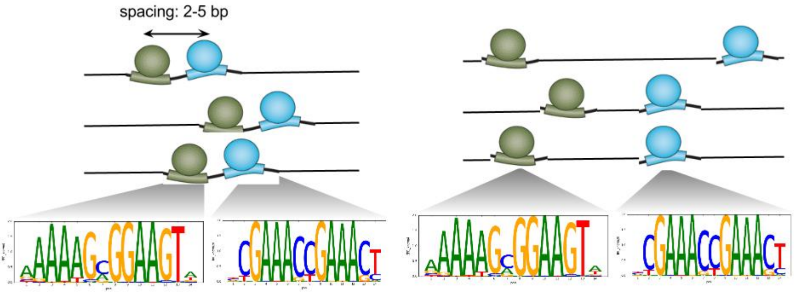


Sequence Simulations

```
print_available_simulations()
```

Simulation Name	"Positive" class sequence	"Negative" class sequence
simulate_single_motif_detection		
simulate_motif_counting		
simulate_motif_density_localization		
simulate_multi_motif_embedding		
simulate_differential_accessibility		
simulate_heterodimer_grammar		

Heterodimer sequence simulations



Positive class of genomic sequences containing two motifs with relatively **fixed spacing**

Negative class of genomic sequences containing two motifs with **random and variable spacing**

Representation of motifs (patterns)

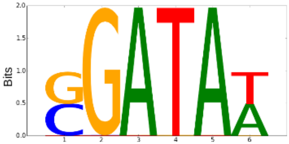
GGATAA
CGATAA
CGATAT
GGATAT

Set of aligned sequences
Bound by TF

$p_i(x_i = a_i)$

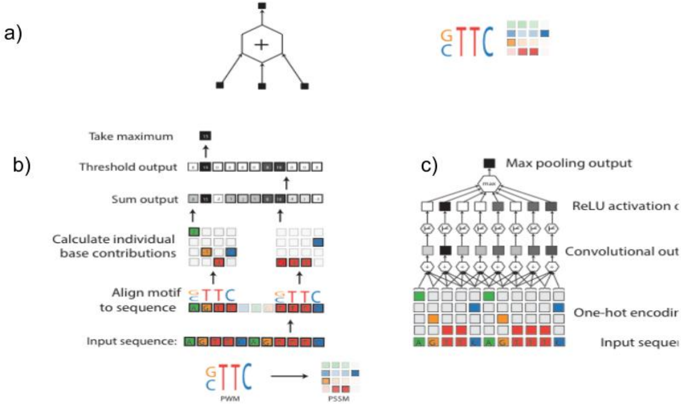
A	0	0	1	0	1	0.5
C	0.5	0	0	0	0	0
G	0.5	1	0	0	0	0
T	0	0	0	1	0	0.5

Position weight matrix (PWM)

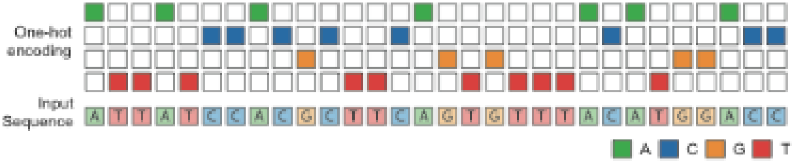


PWM logo

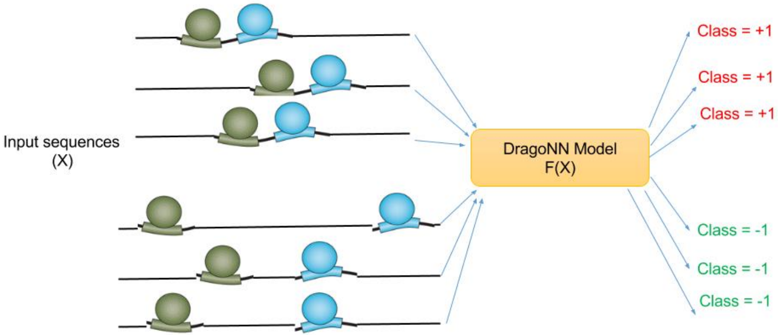
Representation of motifs (patterns)



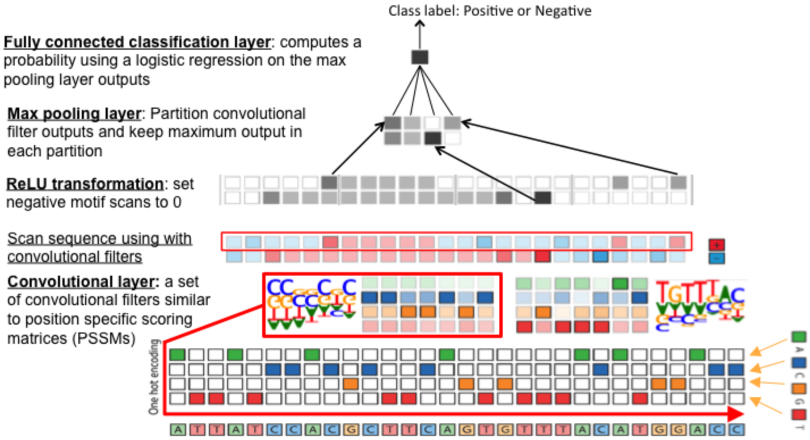
One Hot Encoding



Classification with DragoNN



DragoNN Model



Fully connected classification layer: computes a probability using a logistic regression on the max pooling layer outputs

Max pooling layer: Partition convolutional filter outputs and keep maximum output in each partition

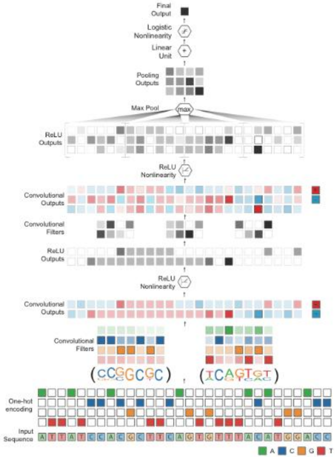
ReLU transformation: set negative motif scans to 0

Scan sequence using with convolutional filters

Convolutional layer: a set of convolutional filters similar to position specific scoring matrices (PSSMs)

Class label: Positive or Negative
+
-
One hot encoding
A
C
G
T
A T T A T C C A C G C T T C A G T G T T T A C A T G G A C C

DragoNN Model



Thank You.

you can follow me through:

www.linkedin.com/in/barathiganeshhb